To appear in *Connection Science*
Vol. 00, No. 00, May 2016, 1–7

The Meaning of the EPSRC Principles of Robotics

Joanna J. Bryson[a]*

[a] *Department of Computer Science, University of Bath, Bath, UK*

(*v1.0 released May 2016*)

Keywords: EPOR, agents, ethics, roboethics, robotics, law, policy

Index to information contained in this article

1. Introduction
2. The principles as policy
3. The principle of killing
4. The principle of compliance
5. The principle of commoditisation
6. The principle of transparency
7. The principle of responsibility
8. Conclusion

1. Introduction

In revisiting the Principles of Robotics — as we do in this special issue — it is important to carefully consider their full meaning. The meaning of communication depends both on the context in which that communication is embedded and the author's intent. Here I address first the meaning of the document as a whole, then of its constituent parts. I address this as one of the experts invited to the EPSRC robot ethics retreat, and present at the Principles' conception.

The EPSRC Principles of Robotics (Boden et al., 2011) were generated as a mechanism of impact by a group assembled with little guidance and no set deliverable required. The original intention of the EPSRC robotics retreat seems to have been only the consultation itself, or perhaps even only the fact that an expert consultation occurred. The academics present wanted something substantial to show for their time spent, capturing and consolidating our efforts, discussions, and debates. Our motivation to really achieve some progress for our country and discipline was such that a substantial amount of time of all those present on the final day went into the creation of the three versions of the principles and parts of their documentation. Some of the documentation was extended for the Web page after the meeting, again with at least an effort at consensus by email.

It is right and fitting that there should be a way to examine the Principles, and to the extent that they are policy rather than just a historical documentation of one AI ethics meeting, even update or maintain them. Even national constitutions ordinarily have means for maintenance. However, it is critical to the efficacy of policy documents that they are not easy to change. Policy provides context within which we not only select

*Corresponding author. Email: j.j.bryson@bath.ac.uk

actions but build organisations and products, and as such should be a rudder to prevent dithering. This implies that part of their value lies in their stability, so ordinarily policy must be more difficult to alter than it is to instantiate in the first place. Note that some countries and other political unions have not found it easy to create even their initial constitutions for this very reason.

Note the framing of the previous paragraph. I am not claiming that the Principles are necessarily a policy document, only that they might be. I *am* claiming that they may be only—or else to some extent—a historical document communicating the intellectual agreement of a set of academics in late 2010. In that case, it makes no sense to revise the Principles, only to study them. But we framed the Principles with the hope they would become policy, and to some extent 2016 has seen that hope validated. Not only have the Principles themselves gained more standing internationally (as witness this issue and the meeting that preceeded it), but also they have influenced two works that are closer to being conventional policy documents. The first is now published, the British Standard Institutes "Robots and robotic devices: Guide to the ethical design and application of robots and robotic systems" ($AMT/-/2$, 2016), and the second, now released as a draft for public consultation, is the IEEE's "Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artifcial Intelligence and Autonomous Systems" (The IEEE & Systems, 2016). Both take the Principles' lead to focus on ethical *design* of AI, rather than on ethics *for* AI; that is, all three agree with the Principles position that ethics are a mechanism by which humans govern human society, and AI as a product of our design is our responsibility.

Determining the appropriate status and therefore treatment of the Principles motivates the examination here of their provenance and meaning here. Note further that this article is not just about the Principles, but more generally about how academic interest and investment can be directed and optimised for societal impact. I begin by examining the meaning of 'policy', and the extent to which the Principles conform to that category.

2. The Principles as policy

Technology policy, and policy more generally, is a surprisingly amorphous thing. Like other aspects of natural intelligence, policy is not always found residing coherent in one place. It is not necessarily in the law or even in governance. Much of policy is unwritten and even not explicitly known. One of the great strengths of British culture its innovation of the common law, which acknowledges this and the importance of culture and precedent (Mahoney, 2001). Nonetheless, in the cold light of British government surveillance of academic impact, we have to ask, are the Principles policy? If that question is Boolean, then I think the answer is "yes". They are a set of guidelines agreed by a substantial if perhaps arbitrary fraction of the community they affect, and they are published on government web pages. More importantly, they alter behaviour.

All policy has three components: allocative, distributive, and stabilising (Landau, 2016). The *allocative* is the process of determining what problems are worth spending time and other resources on. In the case of the Principles, this was instigated by the EPSRC (or some organisation above them) out of concern that the British public might reject robotics as they had genetically modified food. We were told the rejection of robotics was perceived as a significant threat to the British economy [1]. Note also that in addition to the government investment in the Robotics retreat, each of the partici-

[1] In 2013, Willetts (n.d.) named big data, and robotics and autonomous systems, as two of eight "great technologies" key to the UK economy.

pants (at least those not specifically paid to attend) also made individual investments, allocating time to the problem of robot ethics, though of course for many this investment was confounded with an opportunity to get better known by their primary funding organisation.

The *stabilising* component is the one that ensures that the policy, once set, is incorporated into society in such a way that it is unlikely either to be quickly undone or to become much of a liability or matter of controversy. In the case of the Principles this has evidently been achieved at least to some level since we are celebrating their sixth anniversary with this special issue. From talking to other authors, I know of none of us entirely enamoured with the final product, but all respect the (representative) democratic process by which they were achieved, and the importance of their colleagues' mutual commitment to the final product. I have been very happy to see the Principles further reified into formal policy as mentioned in the Introduction. I am also working with both academics and governments to promote at least some of their implications into law, and I know others of the authors are doing the same. The process by which such things are accomplished is obscure, reliant as it is on political fortune and social connections, but good evidence and reason can still be of benefit. Many politicians and their researchers really do want to be effective.

I leave for last the most controversial aspect of policy: the *distributive*. At its base, all policy is about action selection, and that implies the allocation or rather reallocation of resources. Politics tries to brush over this, since it necessarily goes against the grain of those from whom the resources are reallocated, even in the cases where those individuals stand to gain net benefit. We hate to lose control, but policies are *for* control. "Tries to brush over" is in fact an understatement; making redistribution palatable may just be the core project of politicians.

In the case of the Principles, the government had very specific concerns about individuals who had been in the media promoting fear of robots, and were very clear in their desire to find ways to shift media attention and public impressions towards robot safety. However, it was the participants who brought forwards the means of doing this, with the two other major shifts from popular sensationalism to pragmatism evident in the Principles:

- that robots are not responsible parties under the law, and
- that users should not be deceived about their capacities or status.

The Research Council representatives knew this particular redistribution of attention and therefore power and money would anger some of their outstanding funding recipients, and the participants knew the same about some of our colleagues. Nevertheless, there was striking unanimity amongst the academics that the greatest moral hazards of robots was their charismatic nature and the incredible eagerness many people have to invest their own identity in machines' (Bryson & Kime, 2011). This over-identification leads to striking confusion about robots' nature that all of us had witnessed. Such charisma and confusion opens the door for all kinds of manipulations by corporations and governments, where the robots could be set up as responsible—or even as surrogate—for human lives or values.

Now that I have established that the Principles can be viewed at least partly as policy, I turn to the meaning—the context and motivation—of each of the five.

3. The principle of killing

Robots are multi-use tools. Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security.

The first three principles were intended as corrections of Asimov's laws (Hirose, 1991). Robots are not responsible parties, so they could not kill. Instead, robots should ideally not be *usable as tools* for killing. This simple rule made the transfer of moral subjectivity clear, and simultaneously met the pacifist desires of most present. However, pragmatically, robots were already being used as weapons of war. Laws that are unenforceable are generally considered to be of little or even negative utility (McNeilly, 1968). We were persuaded that leading with a principle known to be false would significantly decrease our chances for cultural or policy impact. The meaning of the first principle might therefore seem neutralised by the compromise of the exception, but that robots are not to be weapons in civil society is still an important social point[2]. Beyond this, the fact that practical policy has to take into account the needs of the government to address both security and industry[3] also has meaning. However purely-academic some of us may wish our discipline to be, the fact that many of its products have immediate utility means that we cannot avoid impact on our world.

4. The principle of compliance

Humans, not robots, are responsible agents. Robots should be designed & operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, including privacy.

The second Asimov law has to do with following instructions, but even the notion of obeying implies moral agency (Bryson, 2016). The meaning of this principle as originally drafted was that robots are ordinary technology and conform to ordinary standards and laws. In the shaping of the Principles as a suite, the second principle came to be the one that communicated further some of the peril of AI in general, and AI mistaken for a moral subject in particular. The emphasis on privacy reflects the special concern of a perceiving, intelligent, physical agent occupying the exact same space as a human family. A robot is fundamentally immersed in the human *umwelt*, more than any previous technology or pet, perhaps even more so than some humans in a household such as children. It has access to written and spoken language, social information, observed schedules etc. Further, it may be mistaken for a pet or other trusted family member, its special abilities for perfect communication to the outside world temporarily forgotten, or its abilities to learn regularities and classify stimuli (Kuzuno & Tonami, 2015). In these cases, private information may be unintentionally stored in a public cloud, or even a supposedly private cloud susceptible to hacking. Forcing such a novel, human-like technology into compliance with standard, legal norms of privacy and safety is a non-trivial task

[2]One that was very much in the news in 2016, when the Dallas police department killed an assassin using a teleoperated robot. Because unlike the military, police are only authorised to take life when in direct danger, remote killing (including bombing) by police is generally considered illegal. However, in this case no charges were brought against the police officers involved.

[3]As of 2014, the UK is the world's fifth largest arms dealing nation (Institute, 2015).

5. The principle of commoditisation

Robots are products. They should be designed using processes which assure their safety and security.

The final Asimov law is self protection, but robots have no selves. Instead this principle focusses on protecting humans from robots at the level of the robot's basic soundness. The principle again brings us into awareness of the non-special manufactured nature of the robot, in an attempt to head off avoidance of legal liability by claiming robots have a unique nature. The manufacturer of a robot should have exactly as much responsibility for the machinery working to specification as the manufacturer of a car or a power tool. In fact, robots might be cars or power tools, but if so they should be more rather than less safe than the conventional variety of either.

Note that this principle draws attention to the real topic of the Robotics retreat. We were not discussing what robotics could possibly be. We were discussing what sort of commercial robot products should be allowed to be manufactured, sold, or operated in the United Kingdom. Many have mistakenly asserted that we neglected the possibility of robot sentience or moral patiency (e.g. Gunkel, 2012; Prescott, 2017). This is wrong. The Principles are agnostic to that possibility, but specify that commercial products that will be sold and owned in the U.K. should not be built to be persons (cf. Bryson, 2009).

6. The principle of transparency

Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.

The first three principles established the legal framework for the manufacture and sales of robotics as being identical to other products. The last two are intended to ensure that this status is also communicated to the user. The principle of transparency seeks to ensure that individuals do not overinvest in their technology, for example hiring a house sitter to keep their robot from being lonely (cf. Bryson, 2000).

Some roboticists object to this principle because deception is necessary for the efficacy of their intended application, such as making people to not feel lonely so they are less depressed. Others contend that this principle denies the possibility that robots should be more than ordinary machines. The first argument is empirical, open to experiment. First it needs to be established that there is no way to trigger emotional engagement without deception. Given the extent of emotional engagement that is established with fictional characters and clearly non-cognizant objects, this seems unlikely. If a requirement for deception is however experimentally established, then and only then can any tradeoff between the costs and benefits of such deception be debated.

The second argument however is incontrovertible. The authorship we have over artefacts is a fundamental part of their machine nature—AI is definitionally an artefact. We might even argue that this principle is self-limiting. In the unlikely event the application of AI did alter fundamentally what it means to be a machine, then communicating this modified machine nature would still meet, in fact be required by, this principle.

7. The principle of legal responsibility

The person with legal responsibility for a robot should be attributed.

Finally, the fifth principle communicates robots' status as artefacts in the most fun-

damental way possible. They are owned, and that ownership must be legally attributed. The fact that robots are constructed and owned is the reason Bryson (2009) argues that we are ethically obliged not to design or construct them to be psychologically or morally persons — because owning persons has been established by most modern societies to be unethical. The argument here is not that there exists person-like robots that we should demote in status legally, but rather that the necessarily-demoted legal status means that we should not cause personality to be a feature of any robot legally manufactured (Solaiman, 2016).

However, the Principles of robotics do not go to this extreme of futurism. As I said earlier, they focus on communicating the present reality to a population so eager to own and identify with the super-human that they might easily be lead to believe that a robot badly manufactured or operated is itself to blame for the damage inflicted with it. If you hear a horrible noise and find a car smashed into your house, you can quickly and easily identify the owner of the car, even if the car is presently empty, simply through its number plates or in the worst case through serial numbers. The idea is that the same should be true if you find a robot embedded in your property or person. The participants in the robotics retreat accurately predicted a problem now already present in our society because of drones, and one that is now being addressed in some nations with mandatory licensing such as the committee recommended.

8. Conclusion

To summarise, the EPSRC principles are of value because they represent at least to some measure a sensible policy having efficacy and constructed at significant taxpayer and personal cost. While no policy is perfect, ideally any policy should only be replaced by a new policy with an equivalently high or higher level of investment both by government and domain experts. This is necessary to avoid dithering, or in the terms of policy, for stability. The purpose of the Principles is to provide consumer and citizen confidence in robotics as a trustworthy technology fit to become pervasive in our society. The individual principles each represent substantial concerns of the experts and stakeholders, though sometimes that representation is itself not perfectly transparent. The overall goal was to clearly communicate that responsibility for safe and reliable manufacturing and operation of robots was no different than for any other objects manufactured and sold in the UK, and therefore that the existing laws of the land should be adequate to cover both consumers and manufacturers.

It is important to realise that this is not the case for all conceivable robots. It is easy to conceive of unique works of art that qualify as robots and yet are not like commoditised products, or to conceive of robots that are simply built in an unsafe or irresponsible manner. What people have more trouble conceptualising is that there may be cognitive properties such as suffering that might possibly be feasible to incorporate into a robot, but that to do so would be as unethical as putting faulty brakes on a car. The Principles of Robotics do not seek to determine what is possible; they seek to communicate advisable practices for integrating autonomous robotics into the law for the land.

References

AMT/ − /2, C. (2016). *Robots and robotic devices. guide to the ethical design and application of robots and robotic systems* (Tech. Rep. No. BS 8611:2016). British

Standards Institute.

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., ... Winfield, A. (2011, April). *Principles of robotics.* The United Kingdom's Engineering and Physical Sciences Research Council (EPSRC). (web publication)

Bryson, J. J. (2000). A proposal for the Humanoid Agent-builders League (HAL). In J. Barnden (Ed.), *AISB'00 symposium on artificial intelligence, ethics and (quasi-)human rights* (pp. 1–6).

Bryson, J. J. (2009, November). Building persons is a choice. *Erwägen Wissen Ethik*, *20*(2), 195–197. (commentary on Anne Foerst, *Robots and Theology*)

Bryson, J. J. (2016). Patiency is not a virtue: AI and the design of ethical systems. In *Ethical and moral considerations in nonhuman agents.* Stanford.

Bryson, J. J., & Kime, P. P. (2011, July). Just an artifact: Why machines are perceived as moral agents. In *Proceedings of the* 22^{nd} *international joint conference on artificial intelligence* (pp. 1641–1646). Barcelona: Morgan Kaufmann.

Gunkel, D. J. (2012). *The machine question: Critical perspectives on AI, robots, and ethics.* The MIT Press.

Hirose, S. (1991). A dispute over robots. robots of the future. *Advanced Robotics*, *6*(2), 231–241. Retrieved from `http://dx.doi.org/10.1163/156855392X00060` doi:

Institute, S. I. P. R. (2015, February 17). *Top list tiv tables.* (http://armstrade.sipri.org/armstrade/page/toplist.php)

Kuzuno, H., & Tonami, S. (2015). Detection of sensitive information leakage in android applications using signature generation. *International Journal of Space-Based and Situated Computing*, *5*(1), 53–62. Retrieved from `http://www.inderscienceonline.com/doi/abs/10.1504/IJSSC.2015.067998` doi:

Landau, J.-P. (2016, February 17). *Populism and debt: Is Europe different from the U.S.?* (Talk at the Princeton Woodrow Wilson School, and in preparation)

Mahoney, P. G. (2001). The common law and economic growth: Hayek might be right. *The Journal of Legal Studies*, *30*(2), 503–525.

McNeilly, F. S. (1968). The enforceability of law. *Noûs*, *2*(1), 47–64. Retrieved from `http://www.jstor.org/stable/2214413`

Prescott, T. J. (2017). this issue.

Solaiman, S. M. (2016). Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy. *Artificial Intelligence and Law*, 1–25. Retrieved from `http://dx.doi.org/10.1007/s10506-016-9192-3` doi:

The IEEE, G. I. f. E. C. i. A. I., & Systems, A. (2016, December). *Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems* (Tech. Rep. No. 1). IEEE. (Version 1 - For Public Discussion)

Willetts, T. R. H. D. (n.d.). *Eight great technologies.* The Department for Business, Innovation and Skills for the 2010 to 2015 Conservative and Liberal Democrat coalition government. (Speech delivered on 24 January, also a whitepaper)